EuXFEL – online & offline data processing and storage

Martin Gasthuber, Stefan Dietrich, Janusz Malka – DESY/IT
Kryzsztof Wrona, Janusz Szuba - EuXFEL
CHEP16 – San Francisco

HELMHOLTZ ASSOCIATION

LSDMA

European XFEL

DESY

European XFEL - a leading new research facility

The **European XFEL** (**X**-Ray **F**ree-**E**lectron **L**aser) is a research facility under construction which will use high intensity X-ray light to help scientists better understand the nature of matter.

Schenefeld site at the start of user operation

- Location: Schenefeld and Hamburg, Germany
- User facility with 280 staff (+ 230 from DESY)
- 2017 start of user operation

EuXFEL - participants

- Organized as a non-profit corporation in 2009 with the mission of design, construction, operation, and development of the free-electron laser
- Supported by 11 partner countries
- Germany (federal government, city-state of Hamburg, and state of Schleswig-Holstein) covers 58% of the costs; Russia contributes 27%; each of the other international shareholders 1–3%
- Total budget for construction (including commissioning)
  - 1.22 billion € at 2005 prices
  - 600 M€ contributed in cash, over 550 M€ as in-kind contributions (mainly manufacture of parts for the facility)

Facility overview

Schenefeld

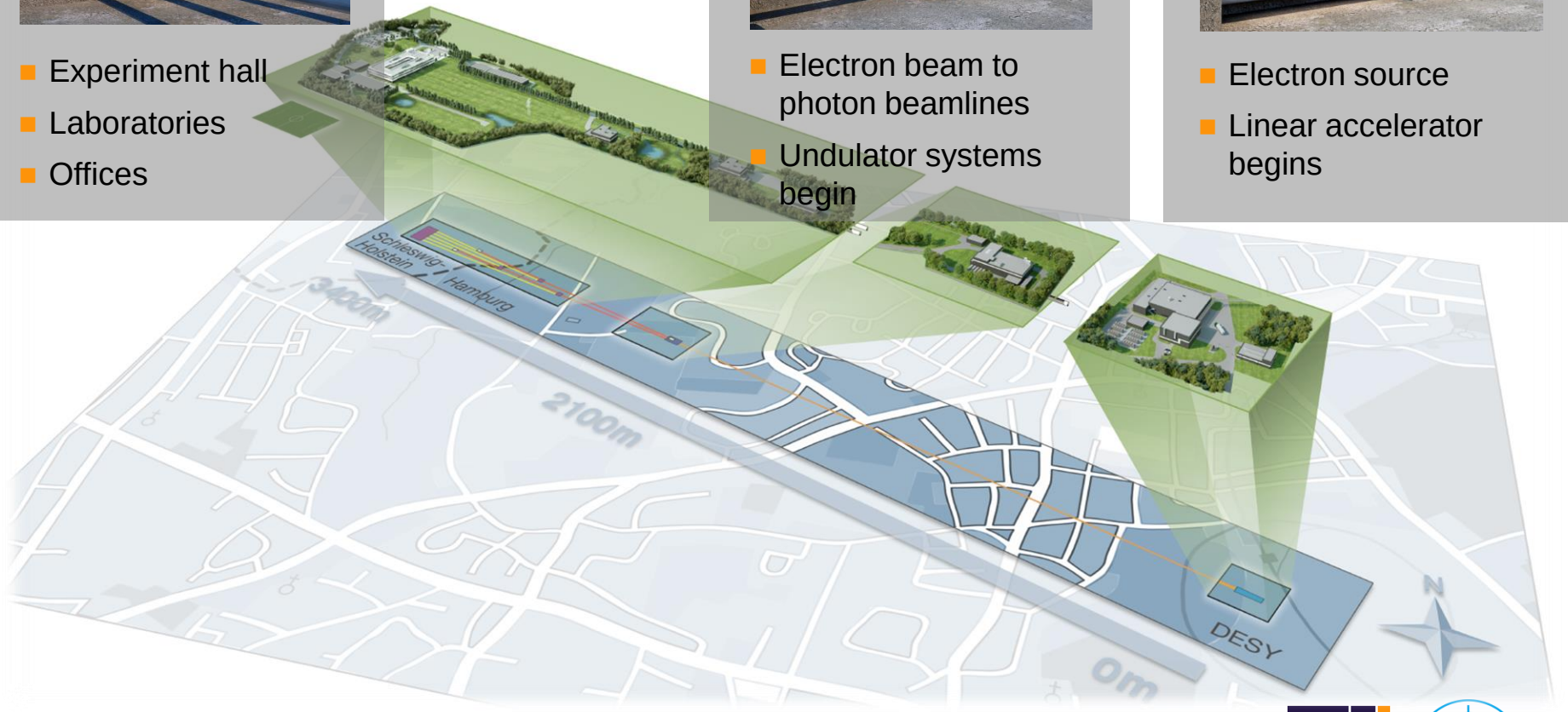
- Experiment hall
- Laboratories
- Offices

Osdorfer Born

- Electron beam to photon beamlines
- Undulator systems begin

DESY-Bahrenfeld

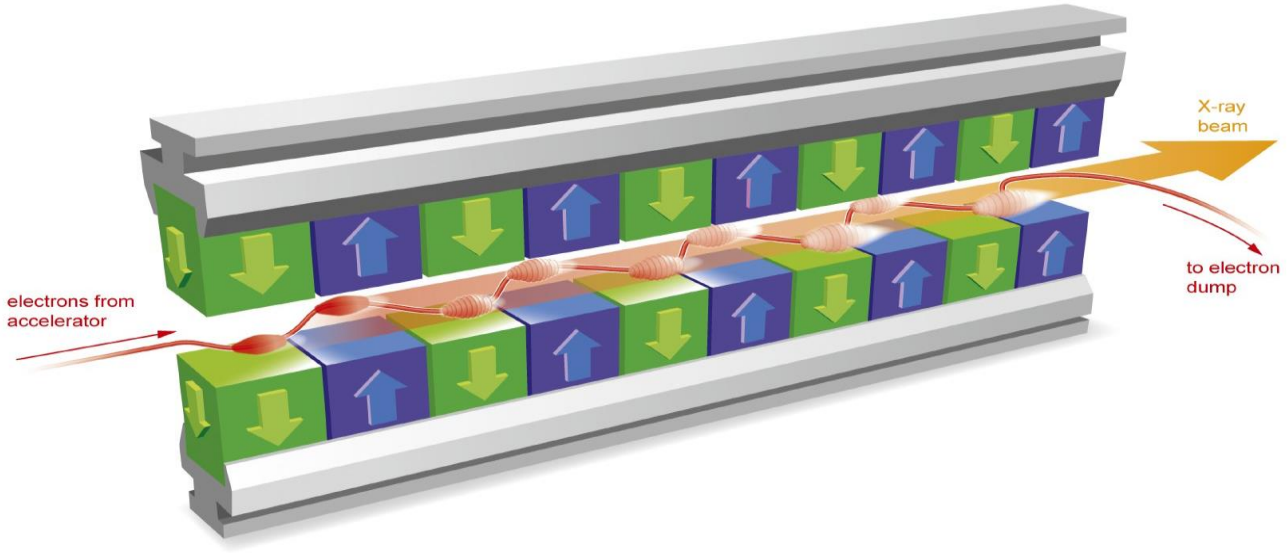
- Electron source
- Linear accelerator begins



European XFEL

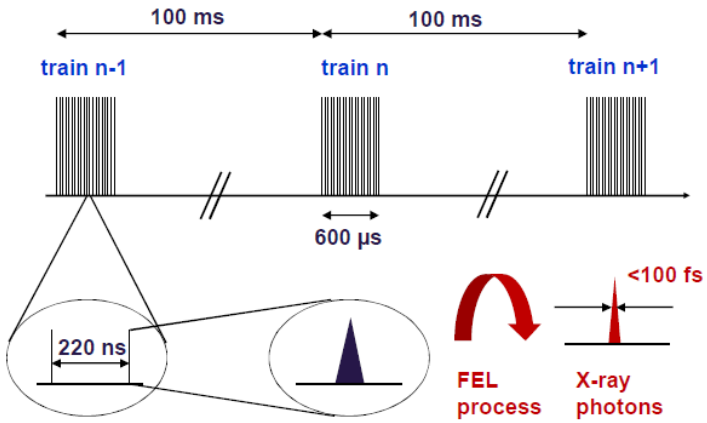
DESY

from electron to coherent x-ray



DAQ Challenges

- Readout rate driven by bunch structure
  - 10 Hz train of pulses
  - 4.5 MHz pulses in train (1-2700 pulses)
- Data volume driven by detector type



Detector type	Sampling	Data/pulse	Data/train	Data/sec
1 channel digitizer	5 GS/s	~2 kB	~6 MB	~60 MB
1 Mpxl 2D camera	4.5 MHz	~2 MB	~1 GB	~10 GB
4 Mpxl 2D camera	4.5 MHz	~8 MB	~3 GB	~30 GB*

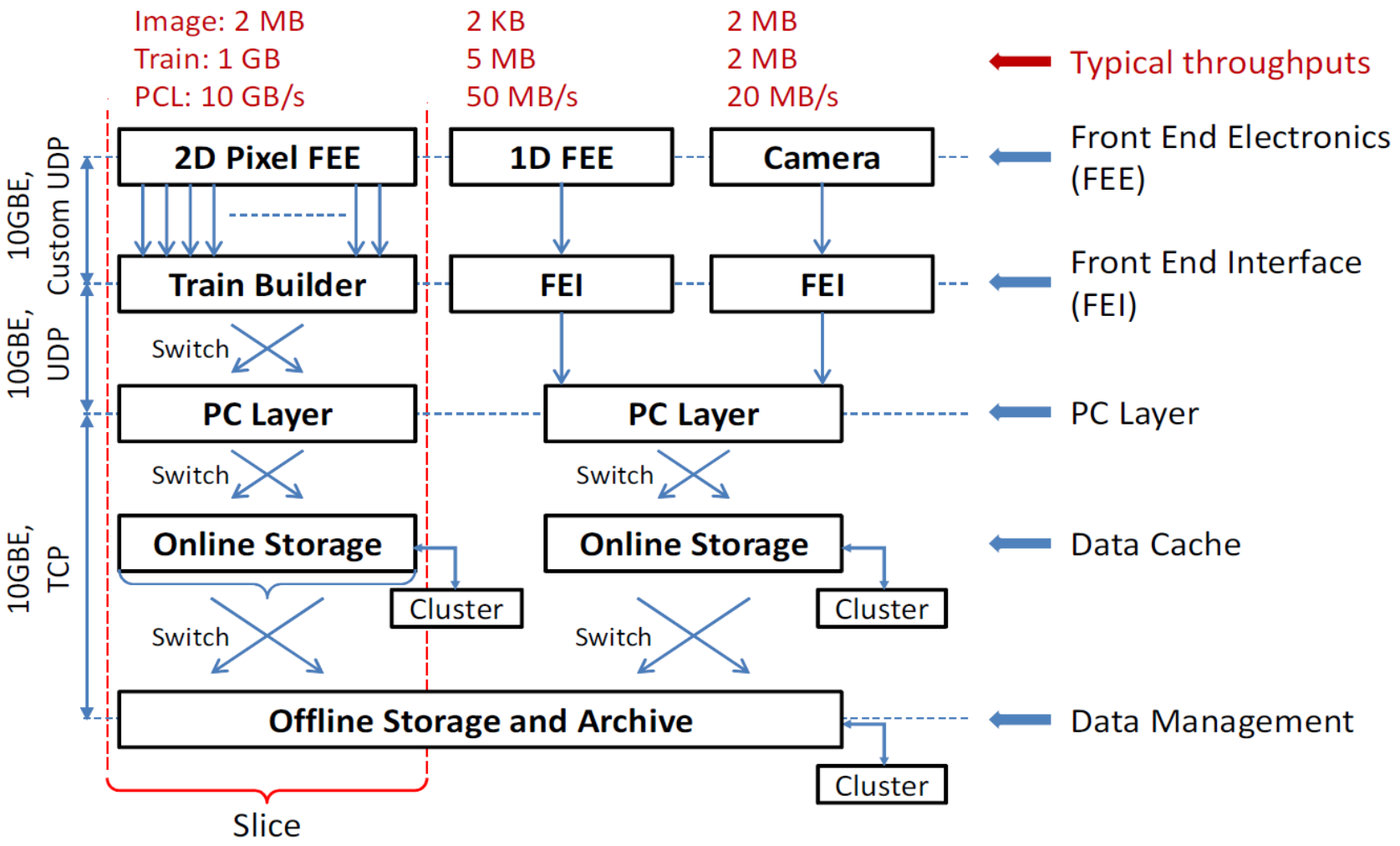
- volume depends on detector type and pulses per train
- 1-N trains per file -> 1GB file or larger

* Limited by AGIPD detector internal pipeline depth (352 img/sec), hence factor 3 compare to LPD 1MPx

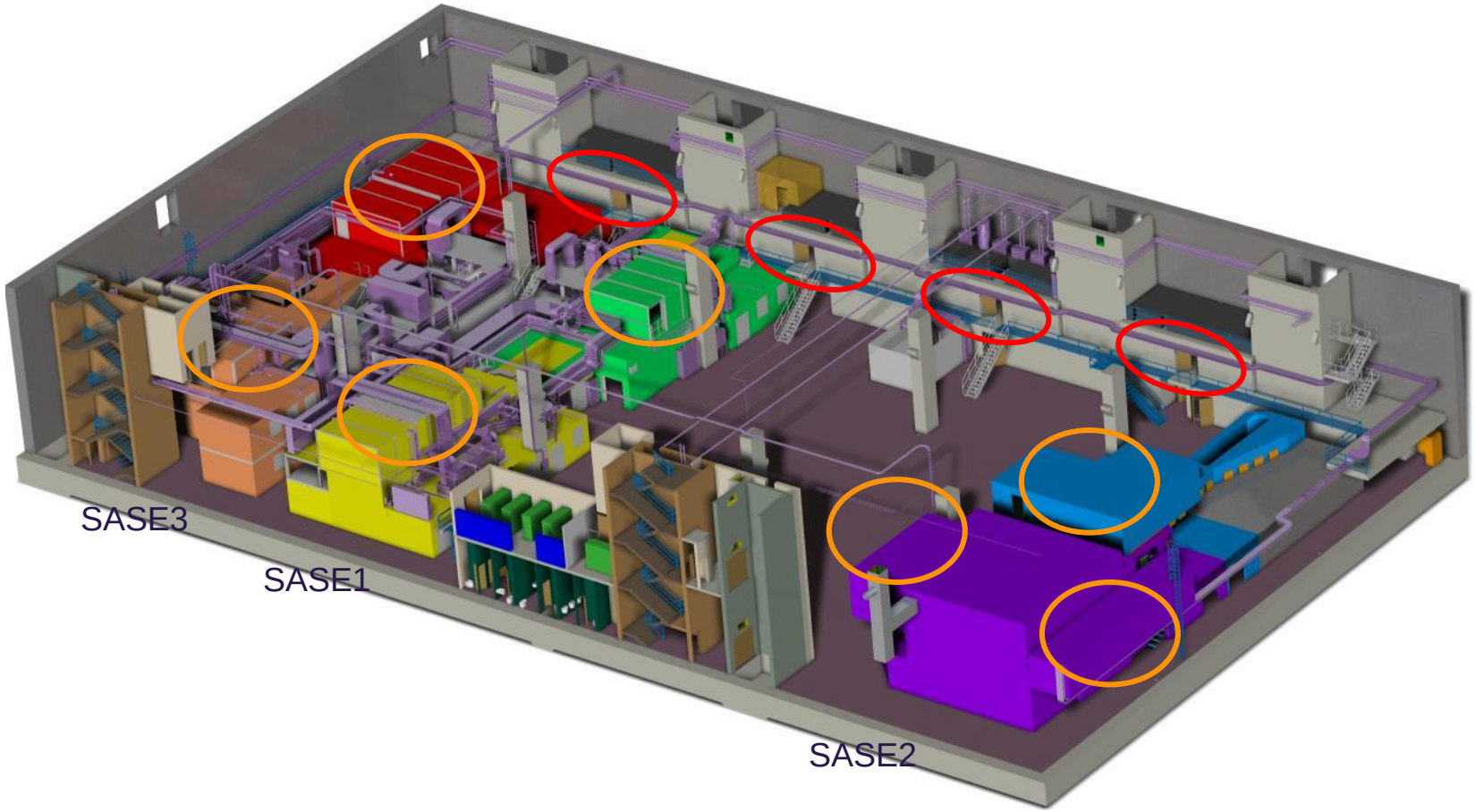
How to cope with that ?

- Standardize detector to DAQ interfaces
  - Multiple 10GE network links to receive data from detector
  - Standard data transfer protocols
  - Standard data formats (HDF5)
- Include software based computing capability into the DAQ chain
  - Data receiving, aggregation, reduction, formatting
  - Enable bad quality data rejection
  - Provide real time overview of collected data e.g. compute statistics, visualize data
- Provide highly optimized infrastructure and resources for data recording close to the experiment station
  - Dedicated network for DAQ
  - Distributed storage systems with controlled/restricted access
  - HPC systems for demanding storage – GPFS on ESS systems

DAQ – data flow and processing



infrastructure locations

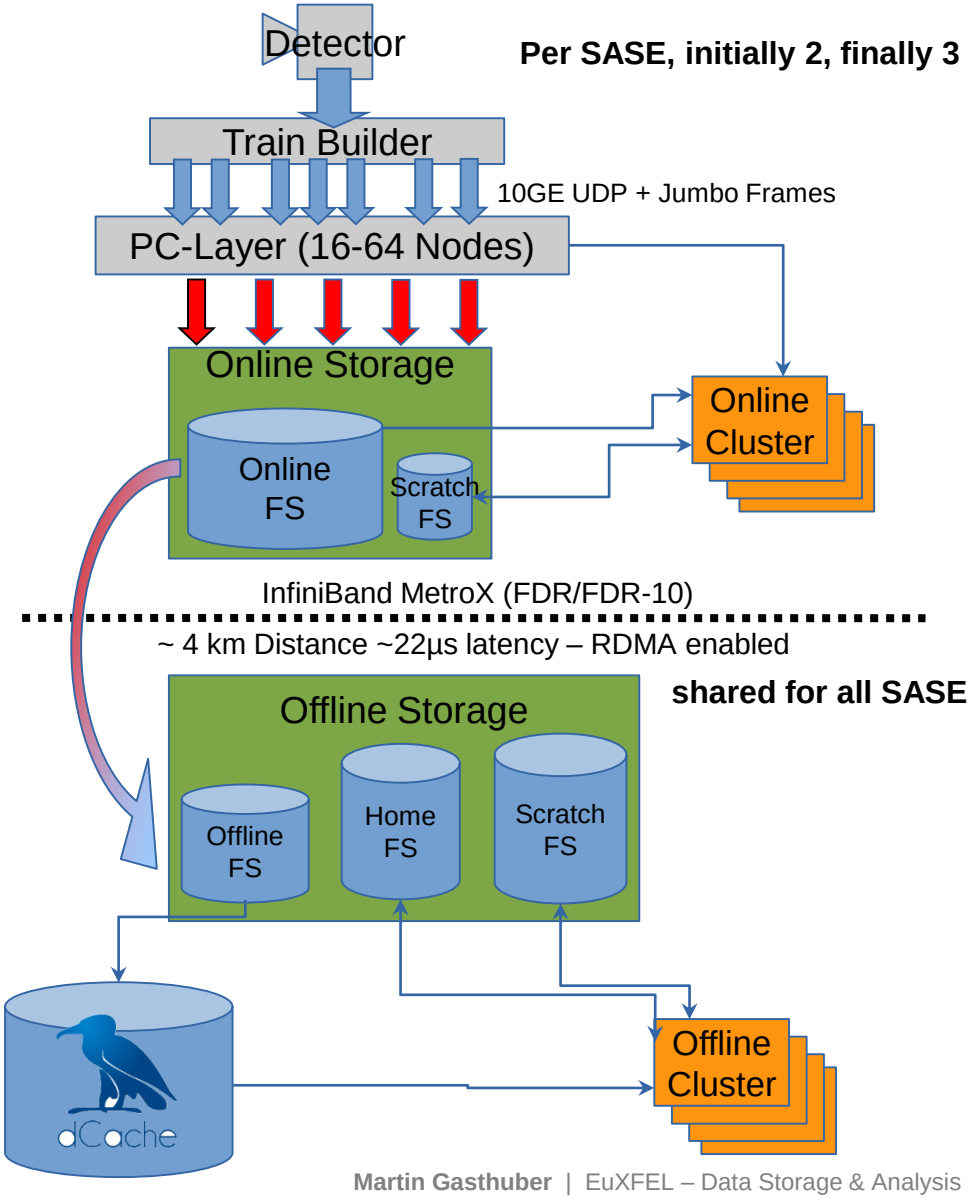


- 4 computer rooms in the experiment hall (red, a.k.a. balcony rooms)
- Dedicated rack rooms for the instruments (orange)

basic things different to HEP

- beamtime (experiment) data taking phase ~week interleaved with second experiment (12 hour shifts) on same beamline (each on its own hutch)
- change active experiment on each shift (i.e. 6 days beamtime leads to 3 days active data taking)
- offline analysis expected to be ~months (2-8)
  - results in ~80-100 offline data analysis running concurrently on shared offline resources
- cold data (sometimes called ‘archive’) alive for ~years (3-8) – active discussion on that – budget constrains might apply. Tape and/or ObjectStores are in consideration
- adopt PaNdata ‘open access’ policy

data flow – more abstract



Train Builder

- Reshuffles picture modules into whole picture
- Pictures shuffled in trains
- Sends single trains per channel

PC-Layer

- Data analysis for monitoring
- Data Reduction, e.g. FPGA based compression
- Veto
- File creation in memory and online filesystem
- every node creates a 1GB HDF5 file every 1.6s

Online Cluster

- 10-80 nodes
- Online data analysis and re-calibration

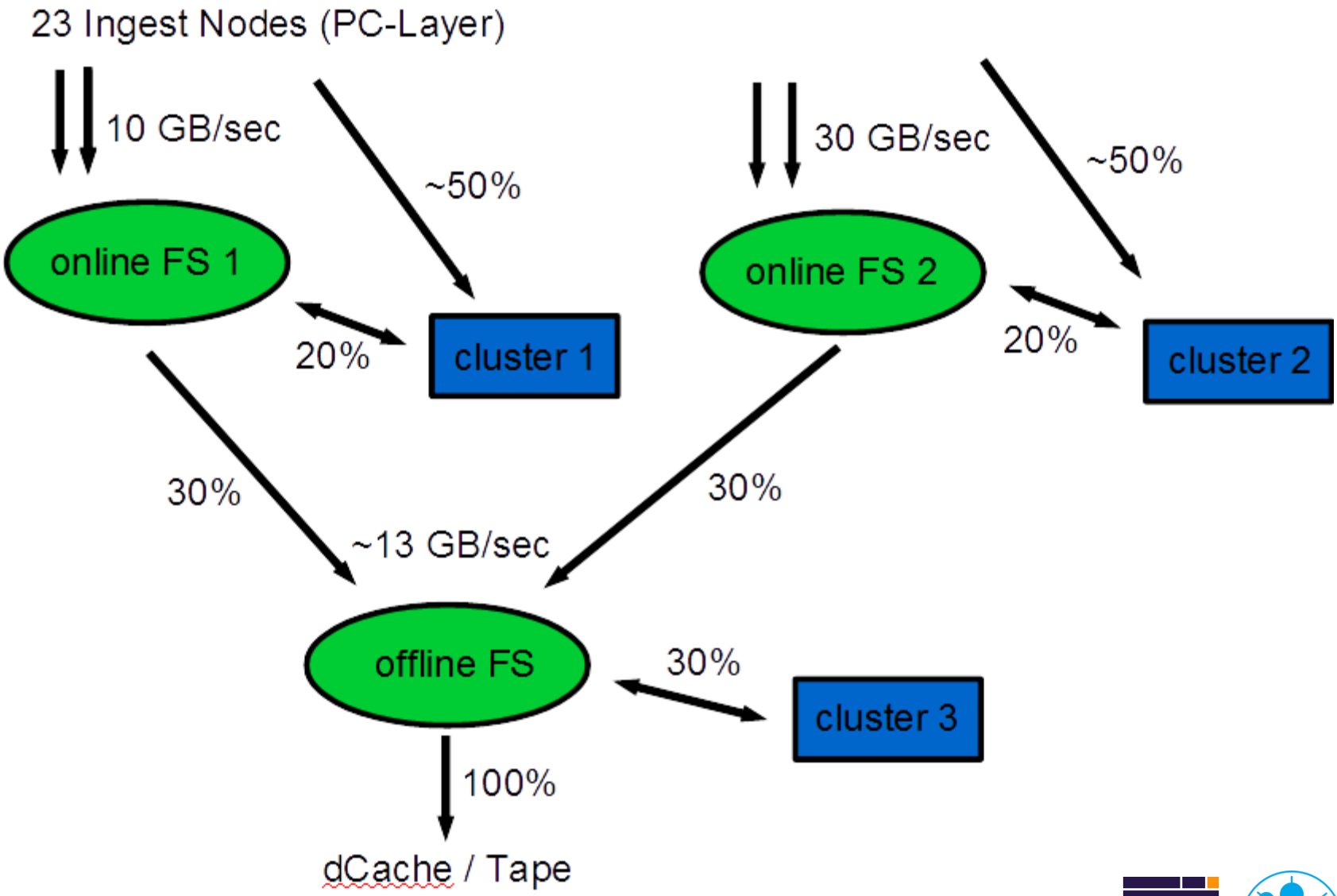
Transfer Online → Offline Storage

- Evaluation: multiple or ‘stretched’ cluster
- Evaluation: GPFS AFM or custom scripts

Offline Storage

- Shared across experimental stations (SASE)
- Data arrives after delay, stored on GPFS
- Copy data to dCache (tape copy, export) ACLs
- Raw data access only from dCache
- Offline cluster stores calibrated data on GPFS
- User analysis from calibrated data

even higher altitude – looking at rates – two beamlines

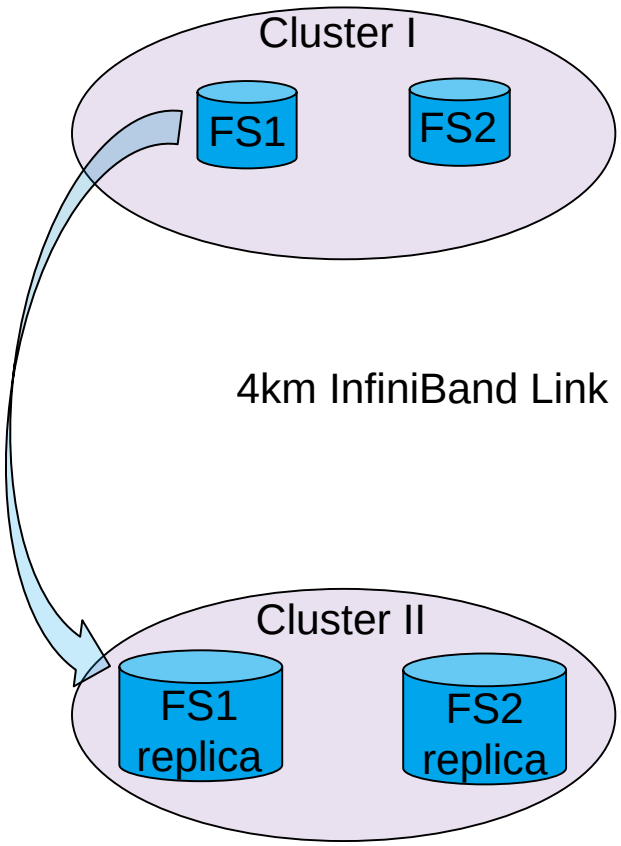


location and phase constraints of major storage instances

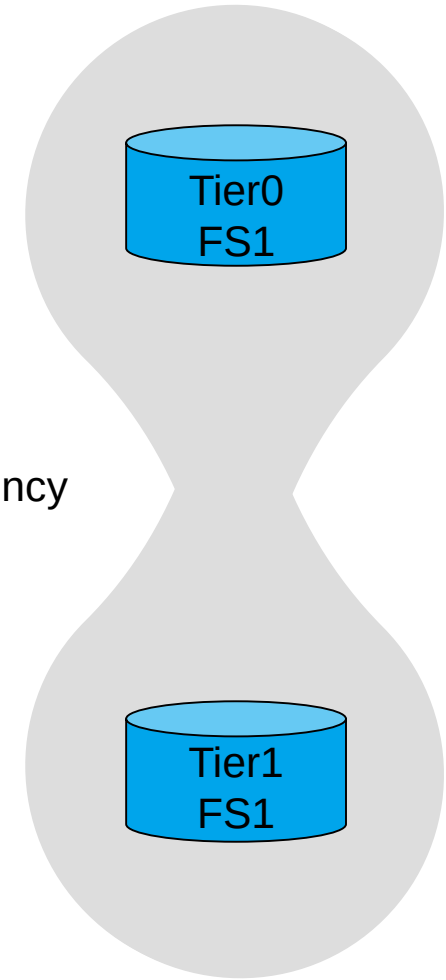
Type	Location	Phase: Commissioning	Phase: Data Taking	Phase: Offline Analysis	Phase: cool down
Shared scratch (accessible online+offline)	online	X	X		
	offline			X	
Raw + Calibrated	online		X		
	offline			X	
Cold (archive)	offline			X	X

GPFS configurations – classic/default, stretch-cluster

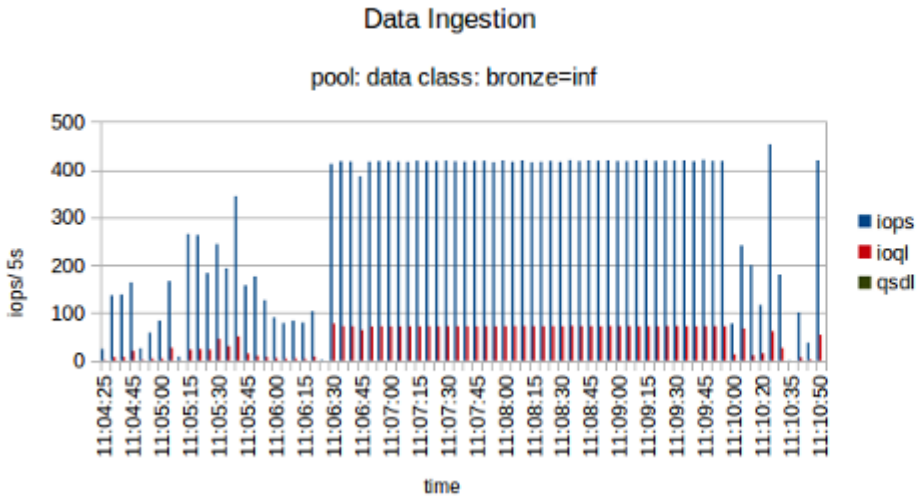
- policy run triggered (scan FS)
- FS event triggered (inotify)



single cluster, single FS, multi-tier

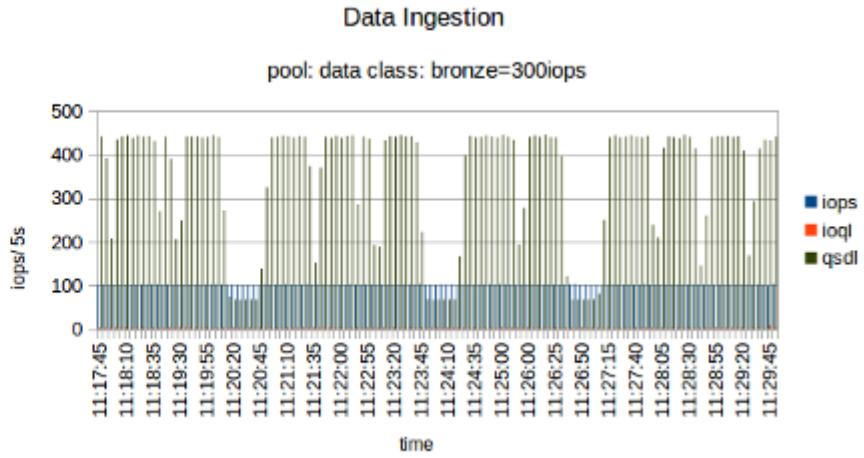


Quality of Service – control IOPS spent profile



no throttling – qsdl == 0
write data (detector simulation)

active throttling – same (IO)
process as above



challenges to continue on…

- bandwidth optimization
  - ingest from detector (the 30GBs detector beamline) – highest priority
  - offline storage to dCache – feed calibration process and ‘copy to tape’ preserving GPFS NFSv4 ACLs
  - control user access – largely non-predictable (QOS)
- prove fault tolerance
  - site failure, link failure (Ethernet/InfiniBand)
- all flash for online storage
  - looks economically feasible (0.5 PB per unit)
  - performance figures under investigation (beta HW/SW)
  - should help a lot to get chaotic user access silently merged
- event driven data migration (instead of policy run)
  - cluster wide inotify – (L)ight (W)eigth (E)vent

status

- roughly 50% of the storage is in order (dCache + GPFS HW) or already in place
- 3000 cores expected by end of 2016 – all IB connected
- EDR/FDR InfiniBand fabric already enlarged (online & offline)
  - ~100 ports per online instance (1/1), ~500 for (single) offline instance (1/2)
- dedicated ESS (GPFS appliance) systems for ‘basic tests’ still being available
  - QOS, LWE, stretch-cluster, object-store integration, weird architectures, …
- all flash testing starts in Nov 2016

protein resources

- from XFEL computing
  - Krzysztof Wrona, Janusz Szuba, Djelloul Boukhelef
- DESY/IT
  - Stefan Dietrich, Janusz Malka, Manuela Kuhn, Uwe Ensslin, Birgit Lewendel, Volker Guelzow, Martin Gasthuber

- more details (only technical) on GPFS being presented by Sven Oehme (IBM/Research Almaden) @HEPiX next week – Tuesday 5pm

- questions ?